(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 143 349 A1

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
10.10.2001 Bulletin 2001/41

(51) Int Cl.^{7}: G06F 17/30

(21) Application number: 00107594.4

(22) Date of filing: 07.04.2000

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU
MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(71) Applicant: IconParc GmbH
80333 München (DE)

(72) Inventor: Brückner, Roland
81371 München (DE)

(74) Representative: Betten & Resch
Postfach 10 02 51
80076 München (DE)

(54) Method and apparatus for generating index data for search engines

(57) The invention relates to a method for generating index data to be provided to a search engine to be used for searching the internet or a non-public network, said index data comprising one or more search indices, said method comprising the steps of: generating at least one search index on a computer based on data stored on or accessible by said computer, said computer being located remotely from said search engine, said index being generated in accordance with one or more settings or selectable options defining which data stored on or accessible by said computer is to be used for generating said at least one index.

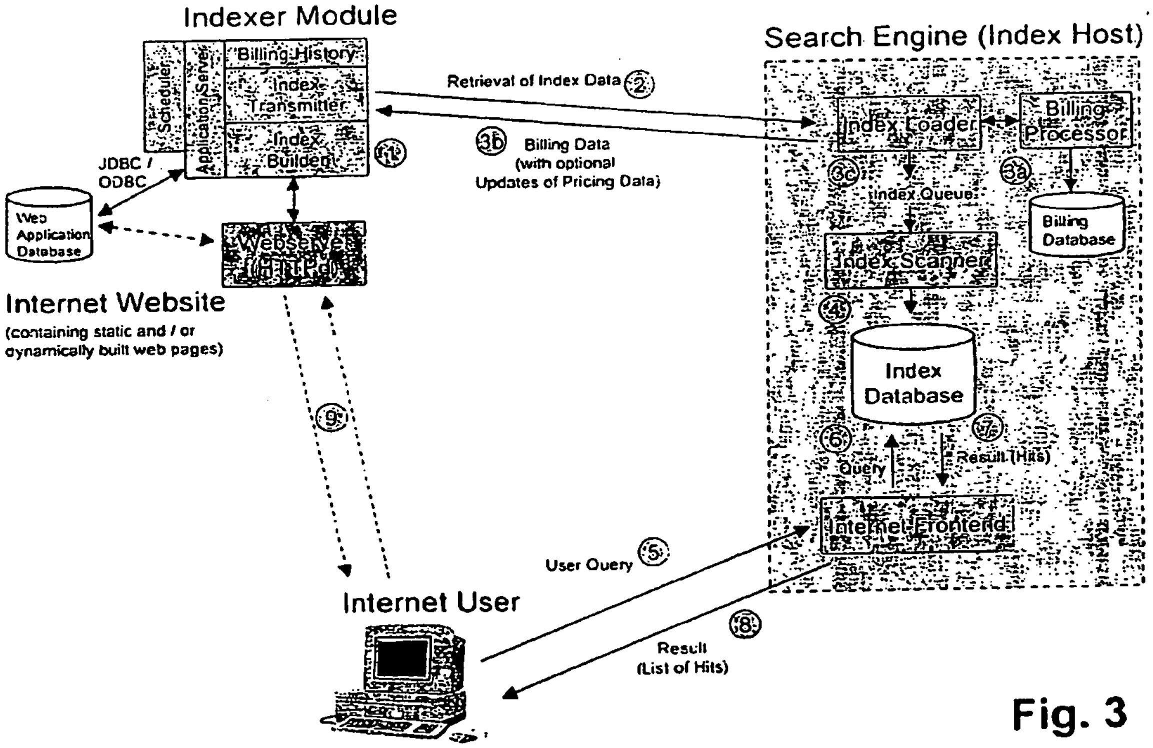

Fig. 3

EP 1 143 349 A1

Printed by Jouve, 75001 PARIS (FR)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 99 12 4674

31-10-2000

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9516971 A	22-06-1995	EP 0734556 A	02-10-1996
		JP 11096243 A	09-04-1999
		JP 10312433 A	24-11-1998
		JP 10312434 A	24-11-1998
		JP 9500470 T	14-01-1997
		US 6049785 A	11-04-2000
		US 5724424 A	03-03-1998
EP 0779587 A	18-06-1997	JP 9167185 A	24-06-1997
		KR 208770 B	15-07-1999
		US 5890137 A	30-03-1999
WO 9907121 A	11-02-1999	AU 8675398 A	22-02-1999
		CN 1267380 T	20-09-2000
		EP 1004086 A	31-05-2000
US 5715314 A	03-02-1998	EP 0803105 A	29-10-1997
		JP 10509543 T	14-09-1998
		WO 9613013 A	02-05-1996
		US 5909492 A	01-06-1999

[0012] Several search engines are capable of maintaining and visualizing the index database's content in a hierarchical (tree-like) catalog structure. Thus, a complementary way of discovering information is available: instead of keying in search terms, a user may -navigate step-by-step through the catalog's (sub)categories (such as sports, finance, technology, news, ...) in order to find relevant pieces of information. Such a catalog's maintenance requires manual effort by the search engine's host even if there is an internet frontend permitting users to add catalog records online (today, no fully automated solution is available on the market). It is crucial to understand that these search engine catalogs only offer very limited functionality in terms of search parameters and datamodel flexibility: at the bottom line it's not a fully-fledged structured datamodel for example for cars or holiday trips but simply a hierarchical navigation model that lets users work their way down the navigation tree starting at more general categories (e.g. "sports") and ending up at more and more specialized categories (e.g. "equipment for river rafting" or "sports events in Atlanta").

[0013] Another approach to store searchable data in a structured catalog on a search engine comprises the capability to handle structured data which is extracted from webpages containing for example data in a tabular manner (e.g. a price list for products offered on a specific webpage). In this scenario pieces of structured data will be transferred from the webserver to the search engine in two different ways:

[0014] The structured data contained in webpages which are meant to be indexed can be copied manually by the staff that runs the search engine. However, this is a time-consuming and error-prone way of maintaining the search engine's database. Thus, it is a common approach to write some sort of "scanner program" that loads and analyzes webpages containing structured data. This must be done on the side of the search engine. That way, the process of retrieving and updating structured data can be automated on a **per-webpage** basis only which is the main catch here: since there are millions of webpages potentially and in practice containing all different formats and models of structured data there is no way of covering even a small share of them as long as it is necessary to write one scanner program for each webpage the structured contents of which is to be analyzed and stored on the search engine.

[0015] All of the conventional approaches, the indexing concept as well as the catalog concepts as known in the prior art suffer from substantial disadvantages.

Summary of the invention

[0016] It is an object of the present invention to provide an improved method and apparatus for generating searchable data to be stored on and made available through search engines.

[0017] According to an aspect of the present invention, at least one search index to be used by a search engine when searching the internet is generated on a computer located remotely from the host on which the search engine is located. This removes the responsibility of generating search(able) index data from the host or operator of the search engine. Instead, this task is performed on a computer where the data which later is to be found through the search engine is located.

[0018] With such a configuration the host, owner or operator of the computer on which the search index data is generated may directly influence the contents of the search index to be maintained by the search engine and thereby he can increase the likelihood that the data he wishes to be found by network users when conducting search operations actually will show up as a query result. A software module hereinafter called an indexer module running on the computer which is hosting the data to be indexed may generate the search index data and this data may then be transferred to the search engine host where it is incorporated into or joins the search (able) index already present there.

[0019] Furthermore such a configuration can prevent inconsistent query results such as that e.g. a link might be broken (i.e. invalid); a link might lead to a web page offering information on other topics than it did when it was scanned by the search engine, etc., since using indexing remote from the search engine the control of index and catalog data is transferred to the information source (= host of an internet website); in other words: the search engine does no longer decide what contents are meant to be gathered.

[0020] Moreover, such a configuration enables automated support not only for static web pages but also for all kinds of dynamically generated web pages since the index generation is carried out on the computer where the dynamic web application is running and not on the search engine host.

[0021] The computer running the indexer module and thus building the index data may be a webserver or any content server being connected to the Internet or to a non-public network; it may also be a normal computer on which a web application is running or on which any internet content can be stored. For example a user of the index generating computer may have his own website being located on another computer such as the server of his internet provider where his webspace is located, but for generating the index he may just transfer his website or web application down to his own computer and may thereon carry out the generation of the index. For that purpose a crawler may be provided in the indexer module for retrieving the internet based content such as a website from the remote server to then build the index data.

[0022] The indexer module may be installed on a webserver as add-on. The thereby built index is then transferred ("pushed") from the originating webserver to the search engine where it joins the already existing index database.

[0037] **Unstructured data** = textual information; e.g. used in conjunction with so-called fulltext searches; when being indexed, unstructured data may be extended to incorporate additional information such as document (webpage) owner, location or timezone;

[0038] **Single index** = index comprising unstructured data; the index may be extended to incorporate additional information such as document (webpage) owner, location or timezone;

[0039] **Structured data** = contents available in a tabular, relational or object-oriented manner; tabular data can be found for example within static or dynamically generated webpages; relational data is typically stored in and retrieved from relational database management systems (RDBMS); object-oriented data is typically stored in and retrieved from object-relational or object-oriented database management systems;

[0040] **Plurality of indices** = collection of indices;

[0041] **Group of indices** = plurality of indices belonging to the same context, i.e. describing a specific model of structured data (e.g. the set of attributes specifying a specific topic such as "Journeys" or "Sports Events" or any other topic)

[0042] In the context of this document index data is generated for use by search engines on the internet or on non-public networks. Thus, index data is a special representation of structured and / or unstructured contents. The index will be used to differentiate between, search for, and locate information resources which are made available through the internet or on a non-public network. Today, each individual resource of information (e.g. a specific webpage) can be addressed by its unique URL (= uniform resource locator).

[0043] **Web resource** = any webpage or document or piece of data accessible through the internet or on a non-public network by means of an URL;

[0044] **Search index** = index data and / or searchable index database maintained by and accessible through a search engine on the internet or on a non-public network; the index database may consist of multiple search indices supporting various datamodels. Each datamodel maps to a specific category of searchable content (e. g. "unstructured text", "IT components", "Books", "Movies", "Tickets");

[0045] **Searchable index** = synonym for search index:

[0046] **Index database** = database storing information which is originally transmitted to the search engine using index data; the database may support unstructured data and / or structured (i.e. tabular, relational, object-oriented) data;

[0047] **Catalog database** = particular characteristic of an index database supporting only structured (i.e. tabular, relational, object-oriented) data in the context of a search engine based on push indexing;

[0048] **Content server** = server system making available various kinds of content resources (e.g. MPEG movies, HTML webpages, Acrobat PDF Documents) on the internet or on a non-public network;

[0049] A first embodiment of the present invention will be described with reference to Fig. 3.

[0050] As soon as the indexer module is installed as a webserver extension it can be used to generate index data for all content that's made available through the webserver. The indexer module reads the webserver configuration and thus determines which virtual paths and virtual servers are available. Index data is then built (1) according to the administrator's settings (selection of static pages and formats, index update schedule, regional classification, etc.). Indexing the contents of dynamically built web pages potentially containing structured data may involve writing program code in a simple script language (or any other suitable programming language supported by the indexer module) to some extent: since structured data typically resides in relational databases the indexer module is capable of connecting to ODBC and JDBC datasources. In conjunction with an easy to learn tag-based programming language any kind of database content may be retrieved and added to the originating index.

[0051] Schematically this can be done by defining the following in a script language:

a) Define data fields (and/or data sources) to be accessed
b) retrieve data from those data fields
c) convert format of retrieved data to match with format required by the index of the search engine
d) define further classification of index data
e) define update interval

[0052] The individual steps are now explained exemplarily in connection with the generation of index data relating to literature. Step a) then defines for example that the data fields author and title and price are to be accessed and indexed. In step b) the data is retrieved. Step c) may define the necessary calculation formula to convert the price from the currency used in the accessed database to the currency used in the index of the search engine. Step d) may additionally be defined that the so generated index data has the regional classification "Germany", which means that a search in the search engine using the regional classification "France" will not lead to the so generated index data being interpreted as a hit. Finally step e) defines the update interval in which the index is to be updated.

[0053] With such an example a bookstore may according to its desire generate index data by accessing its own database and then sending the so generated index data to a search engine host where it is incorporated into the search index of the search engine.

[0054] If step c) and d) are omitted, then the so generated index data may be used in the most general (unstructured) possible index in which each entry only consists of a search term and the corresponding URL. The retrieved data just is indexed and then sent to the search catalog data the search engine's loader module again creates and stores a set of billing data first (3a). Again, billing data is returned to the indexer module which originated the transfer (3b). Then the loader module separates the index data from the catalog data. While the former is placed in the index queue the latter is added to the catalog queue (3c). Due to the different nature of unstructured index data and structured catalog data two separate databases are used. Accordingly, an additional catalog scanner (4b) is required since the index scanner (4a) only handles index data.

[0068] From that moment on, both the index data and the catalog data can be retrieved by internet users who are performing queries accessing this particular search engine's internet frontend with a web browser (5). Queries may be run against only the index database, only the catalog database or both databases simultaneously (6a, 6b). Result sets returned by the databases (7a, 7b) are merged together if necessary. The combined set of results is then transmitted back to the web browser where the query originated from (8). Now the internet user is likely to find the desired pieces of information on one of the web pages the URLs of which are contained in the search results (9).

[0069] Of course it is also possible to generate and transmit catalog data only without generating and transmitting index data.

[0070] In the foregoing embodiments a technology to generate index data remotely from a search engine has been described. This results in several significant advantages over the prior art approach.

[0071] E. g. there is provided an enabling technology for realizing not only a technologically new and inventive approach but also for realizing a new business model: since remote indexing decisively increases the level of quality offered by search engines usage fees can be easily justified; any website host running an indexer module will e.g. be billed according to the number of URLs and / or catalog entries selected for storage on the search engine and their respective update intervals.

[0072] Search engines based on the remote indexing approach are fully compliant to eBusiness applications and information brokerage applications (as opposed to common search engines) since they can easily handle dynamically generated webpages potentially and in practice containing structured and/or unstructured content.

[0073] Furthermore remote indexing is much more bandwidth-efficient than the method employed by common search engines because it does not require transmission of complete web pages between a website and the search engine. Typically, index data shrinks to about 40% of the original page's size. To further reduce the amount of time (and bandwidth) needed to transfer updates index and catalog data may be packed (compressed) prior to transmission.

[0074] Although not expressly mentioned before a catalog host may also be implemented on its own (that is, without the index-handling part dealing with unstructured index data). Today, none of the large, well-known Internet search engines processes structured catalog data alone but this may become a promising approach in the future.

[0075] It can also be imagined that a hybrid configuration of the conventional techology and the technology of the present invention is employed. E.g. a conventionally generated search index may be freely accessed and searched by a user and a search index generated according to the present invention can only be accessed if the user has accepted to be charged for it.

[0076] Communication between indexer modules and the remote indexing search engine may be implemented using a TCP/IP-based protocol. Even though it is imaginable to make use of standard protocols such as HTTP or FTP defining a variation thereof or even introducing a completely new protocol may turn out to be reasonable for applying the present invention.

[0077] The data format used to describe index and catalog data may be in the form of XML or any of its derivatives.

[0078] Apart from regional classification of index or catalog data any other conceivable classification is possible as well.

[0079] It is readily apparent to the expert from the foregoing that hereinbefore there have been mentioned embodiments which are exemplary only and which can be easily modified or supplemented without departing from the spirit and scope of the present invention. E.g. if necessary, index and/or catalog data might be encrypted prior to transmission. Furthermore it should be clear that the elements of the embodiments described above may be realized by means of software, or by means of hardware, or by a combination of both of them.

Claims

1. A method for generating index data to be provided to a search engine to be used for searching the internet or a non-public network, said index data comprising one or more search indices, said method comprising the steps of:

   generating at least one search index on a computer based on data stored on or accessible by said computer, said computer being located remotely from said search engine, said index being generated in accordance with one or more settings or selectable options defining which data stored on or accessible by said computer is to be used for generating said at least one index.

2. A method according to claim 1, further comprising:

   transmitting said index from said remote computer to said search engine to enable the search engine to incorporate the thus transmitted index into one or more of its search indices used for searching

17. A method according to claim 16, further comprising one or more of the followingbilling a search engine user for carrying out a said search method;
    billing the host and/or operator and/or owner of said computer said contents of which is indexed and in turn transmitted to said search engine for making said index data available to network users through said search engine.

18. An apparatus for generating index data to be provided to a search engine to be used for searching the internet or a non-public network using one or more search indices, said apparatus comprising:
    means for generating at least one search index on a computer based on data stored on or accessible by said computer, said computer being located remotely from said search engine, said index being generated in accordance with one or more settings or selectable options defining which data stored on or accessible by said computer is to be used for generating said at least one search index.

19. An apparatus according to claim 18, further comprising:
    means for carrying out a method according to any one of claims 2 to 17.

20. A computer program comprising computer executable instructions for causing a computer to carry out a method according to any of claims 1 to 17.

21. A data structure comprising:
    at least one search index which can be used by a search engine and which has been generated by a method according to one of claims 1 to 17.

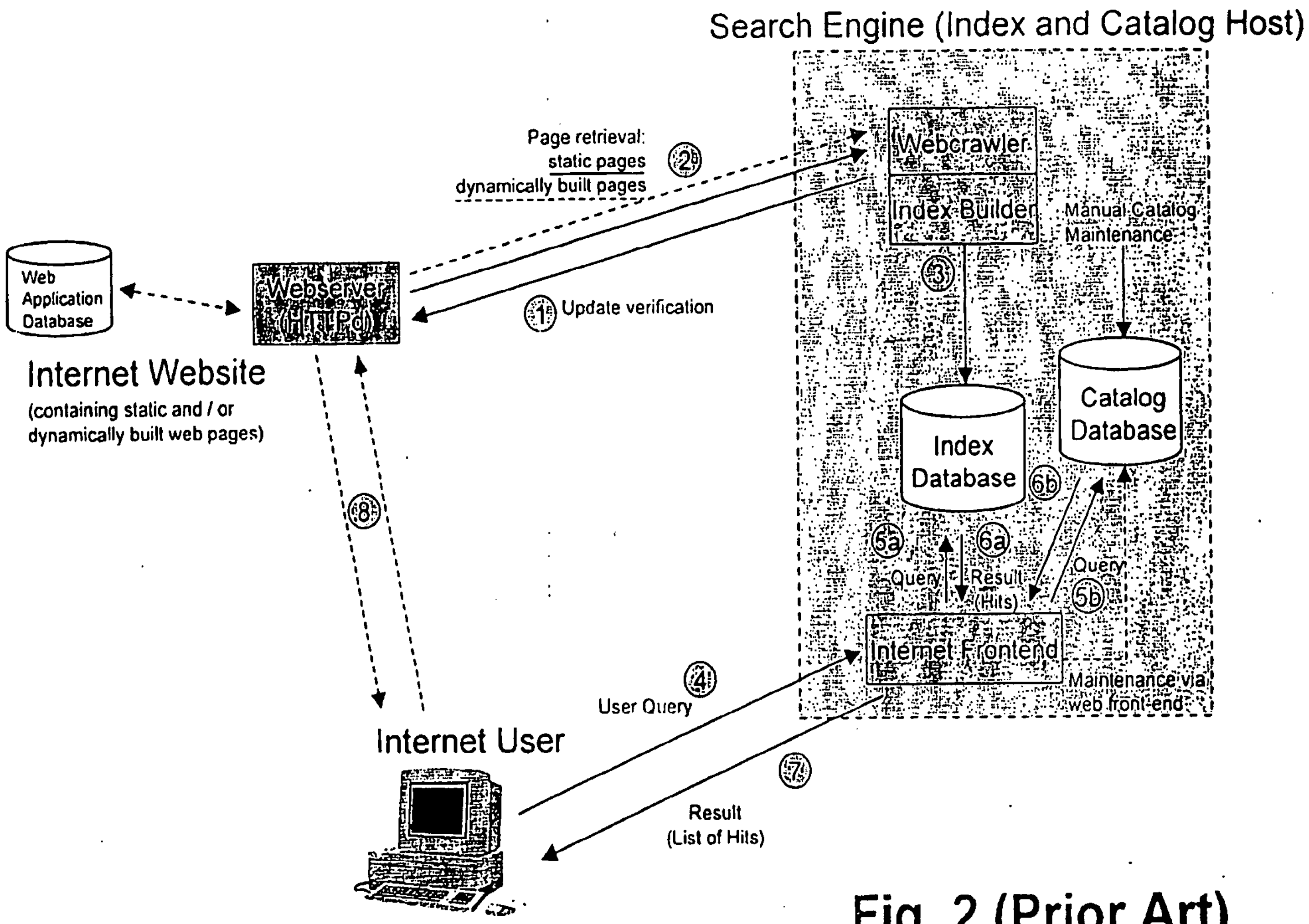

Fig. 2 (Prior Art)

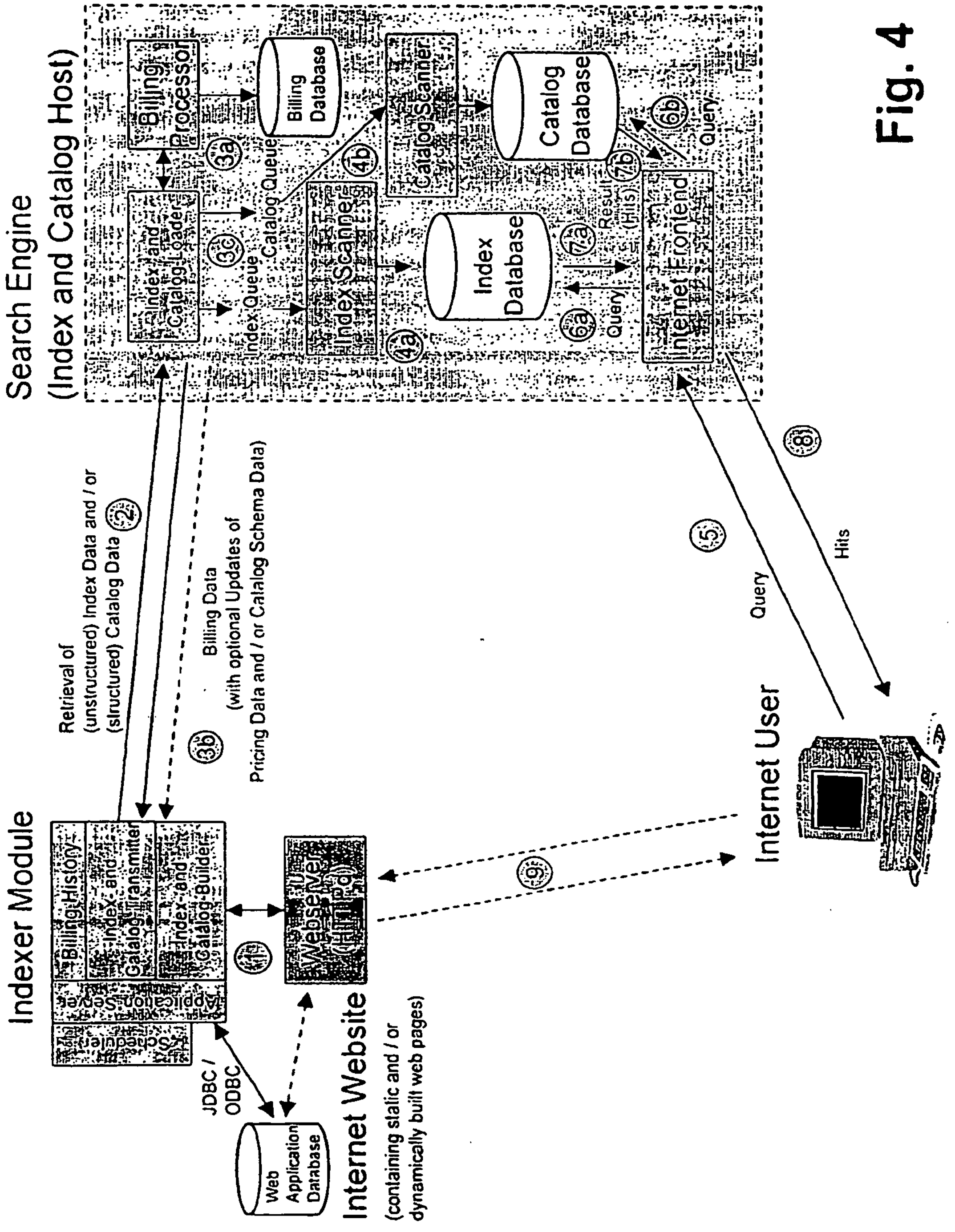
Search Engine
(Index and Catalog Host)
Billing Processor
Billing Database
Index and Catalog Loader
3a
3c
Index Queue
Catalog Queue
Catalog Queue
Index Scanner
Catalog Scanner
4a
4b
Index Database
Catalog Database
6a
Query
7a
Result (Hits)
7b
6b
Query
Internet Frontend
8
Hits
5
Query
Retrieval of
(unstructured) Index Data and / or
(structured) Catalog Data
2
3b
Billing Data
(with optional Updates of
Pricing Data and / or Catalog Schema Data)
Indexer Module
Billing History
Index and Catalog Transmitter
Index and Catalog Builder
Application Server
Scheduler
1
Webserver (HTTPd)
JDBC / ODBC
Web Application Database
Internet Website
(containing static and / or dynamically built web pages)
9
Internet User

Fig. 4

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 00 10 7594

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

06-09-2000

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5905862 A	18-05-1999	NONE	
US 5778367 A	07-07-1998	NONE	

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82